Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons

Briefing Paper for delegates at the Review Conference of the Convention on Certain Conventional Weapons (CCW)

Geneva, 12-16 December 2016

Abstract: Whether the international community adopts the policy principle of meaningful human control or appropriate human judgment, the logical policy and legal conclusions will be the same. That is, if states accept that the obligations of the laws of war fall on human beings to undertake proportionality calculations, precautionary measures and discriminate between military and civilian objects and persons, there are strict limitations the deployment of autonomous weapons systems in time and space due to the requisite information such commanders require to make such judgments. Moreover, the fielding of learning systems may be impermissible because they will frustrate the ability of a commander to know the likely effects of fielding such a system.

Heather M. Roff, Ph.D.
Senior Research Fellow, Department of Politics & International Relations, University of Oxford
Research Scientist, Global Security Initiative, Arizona State University
Future of War & Cybersecurity Fellow
New America Foundation

Introduction

For the past three years, the UN Convention on Certain Conventional Weapons (CCW) has held informal meetings of experts to discuss questions relating to lethal autonomous weapons systems (LAWS). At the most recent informal meeting of experts in April 2016, states agreed to recommend at the CCW's Fifth Review Conference in December 2016 continuing the deliberations by establishing an open-ended Group of Governmental Experts (GGE). The GGE would consider "options" relating to lethal autonomous weapons systems, such as regulation, prohibition or to take no further action.[i]

One of the first questions that a GGE would consider is the definition of an autonomous weapons system. To date, there is some consternation and difference amongst states' understandings. For example, as Canada observed in its 2016 "Food for Thought" paper on the context and complexity of LAWS:

> "Various experts and states have used the term LAWS inconsistently with some suggestion that current technology, including semi-autonomous systems, could be included under this term. Still others seem to be conflating LAWS with existing unmanned systems such as remotely-piloted aircraft."[ii]

Aside from various parties' uses of terms, there appears to be a spectrum of proffered working definitions.

On the one end are accounts that expand the possible sets of weapon systems that fall under the category "autonomous." For instance, Switzerland offered a working definition of AWS as simply:

> "weapons systems that are capable of carrying out tasks governed by IHL [international humanitarian

law] in partial or full replacement of a human in the use of force, notably in the targeting cycle."[iii]

While this definition is more inclusive, particularly as it includes nonlethal systems, its more expansive scope is without prejudice to the question of appropriate regulatory responses. Its intent appears to be to open up space for a case by case analysis of weapons systems, where particular functions or uses may be considered problematic because they would have greater difficulties complying with IHL.

On the other end of the spectrum are those states that seek to define autonomous weapons in such a way to narrow the kinds of systems that would qualify. France, for example, maintains,

"A lethal autonomous weapons system would be characterized by an ability to move freely, to adapt to its environment, and to carry out targeting and launch of a lethal effector (bullet, missile, bomb, etc.). It would operate in complete functional autonomy."[iv]

This definition substantially restricts the kinds of systems classified as "autonomous" as they would require not only the ability to adapt, but to select targets for attack and potentially self-launch.

Similar to Switzerland, the International Committee of the Red Cross (ICRC) focuses on the "critical functions" approach to autonomous weapons, where the concern is over the functions of a system that enable it to select and attack targets without human intervention. The ICRC, moreover, is urging states to "set limits on autonomy in weapons systems to ensure that they can be used in accordance with international humanitarian law (IHL) and within the bounds of what is acceptable under principles of humanity and the dictates of public conscience."[v]

Defining what constitutes an autonomous weapons system, and how it differs from presently fielded automated systems, is not merely a technological concern, it is a political one. As such, it will be up to states to decide.

Nevertheless, this is not to say there is no sense of agreement amongst states at present. Indeed, what has emerged over the past three years is the consensus that states agree they carry an obligation to "respect and ensure respect" for IHL "in all circumstances."[vi] This basic respect for IHL entails that belligerents take the necessary steps to ensure compliance with it. This thin reed of agreement helps to structure some of the worries over autonomy in weapons systems. As the United Nations Institute of Disarmament Research eloquently reminds us, "ultimately the autonomy question is really about what control/oversight […] we expect humans to maintain over the tools of violence that we employ."[vii]

"Meaningful Human Control" and "Appropriate Human Judgment"

Where coalescence between stakeholders has emerged is in broad and general approaches to the need for some form of human control and oversight. Some states have accepted the "meaningful human control" approach, or "effective control" or simply "human control", others prefer "appropriate human judgment."[viii] Whatever the terms ultimately become, there is consensus that no one wants weapons that operate *out of human control*.

Meaningful Human Control

At this year's April CCW Meeting of Experts, Richard Moyes of Article 36 and I published a briefing paper on what we view as encompassing meaningful human

control.[ix] Here, we argued that meaningful human control is best considered as operating at three different layers: *ante bellum*, *in bello* and *post bellum*.[x] At each layer, there are systems, processes, doctrines, laws, and rules designed and enacted to enable human control of, and thus responsibility and accountability for, the use of armed force.

We claimed that "each of these layers also serves to shape and to condition the others, and consideration in all of these areas should inform the next steps for ensuring meaningful human control in the future." In short, meaningful human control requires a holistic approach to the design, acquisition and use of tools of violence.

Furthermore, with regard to use, we argued that human commanders must have *meaningful human control over direct attacks*. This requires a human commander to weigh "her expectations of using a certain technology in a specific context against the risks of unwanted outcomes (while recognizing there are thresholds for accepting certain risks.)" In short, IHL requires human commanders to utilize their judgment when planning or deciding attacks. That is, the law applies to persons and not to things.

Appropriate Human Judgment

However, some still feel that this approach is too general and is ultimately meaning*less*. Others feel that focusing on "control" is too restrictive and prefer that the standard be "appropriate human judgment." This standard would place the emphasis on the human commander or operator and her capacity to judge the likely effect of using an AWS in a particular instance of armed conflict.

For example, the United States Department of Defense Directive 3000.09 states that "autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force." Likewise, Israel has noted that:

"it is safe to assume that human judgment will be an integral part of any process to introduce LAWS, and will be applied throughout the various phases of the research, development, programming, testing, review, approval, and decision to employ them."[xi]

The requirement that these systems be *designed* to allow human judgment generates two subsequent positive obligations:

1. That humans deploying the systems must understand how they will operate in realistic environments so that humans can make informed decisions regarding their use, and

2. To satisfy this obligation, autonomous weapon systems require adequate levels of operational testing, verification, validation and evaluation. This step is required to ensure not only compliance with IHL, but also to provide empirical evidence of system reliability and predictability that informs human decision makers.

Moreover, the most recent US *DoD Law of War Manual* states that:

"6.5.9.3 Law of War Obligations of Distinction and Proportionality Apply to Persons Rather Than the Weapons Themselves. The law of war rules on conducting attacks (such as the rules relating to discrimination and proportionality) ***impose obligations on persons***. These rules do not impose obligations on the weapons themselves; of course, an inanimate object could not assume an "obligation" in any event."[xii]

In short, if one supports the US policy requiring appropriate levels of human judgement and accepts its claim that key law of war obligations apply to persons and not weapons, then very similar conclusions regarding meaningful human control and appropriate human judgment emerge.

Both policy approaches require that a human make proportionality calculations and undertake all feasible measures for precaution in attack. The weapon system cannot be tasked with making a proportionality calculation or estimating whether the principle of precaution is met.

Necessary Limits on Autonomy in Weapons Systems

If we accept the premise that the law of war obligates human commanders to undertake proportionality calculations and ensure precautions are met, then two conclusions follow from this:

1. A human commander cannot meet her obligations if she lacks sufficient information about context *and* the likely or portended effects of using an AWS in a particular situation. She must be able to estimate that the foreseen effects of targeting a legitimate military objective are proportionate in relation to a direct military advantage.

2. Sufficient information about context and likely effects precludes autonomous weapons systems from operating without strict bounds in space and time.

Exploring this point further, we see two further related consequences:

a. Since proportionality, discrimination and precaution require timely and sufficient information, deploying a weapon system over an extended period of time and space *without communications* is likely to violate the human's obligation to undertake a proportionality calculation either because the situation will have changed, thus requiring a new proportionality calculation, or because the human will have *de facto* delegated this task to the weapon.

b. To ensure the appropriate amount of information is provided to a human so that she can fulfill her obligations under IHL, any weapons systems deployed for extended durations need to have a communications link, as well as to communicate back for authorization to use force. To avoid unwanted or unintended engagements then would require humans to maintain adequate situational awareness to authorize the use of force in that situation.[xiii]

Furthermore, it may also be useful to view the necessary limits on the use of AWS from the perspectives of "positive" and "negative" control. Positive control is "the assurance that authoritative instructions to perform military missions will be carried out," and negative control is "the prevention of any unauthorized use."[xiv]

Positive control for an AWS would entail that the systems *work as designed and intended*. They cannot be unreliable or unpredictable, and a human commander must have sufficient knowledge and trust that the system will in fact work as intended.

Negative control for AWS would also require similarly high levels of assurance are met so that systems are not used without authorization. "Authorization," however, would require the commander to undertake all proportionality, discrimination and

precautionary calculations, and to never delegate this task to the weapon.

The positive and negative control framing is useful because it shows us that beyond the technological concerns, such as software design and the weapons platform, there is a requirement for further doctrinal and command systems in place to assure positive control is met. Moreover, negative control does not preclude any use but only unauthorized use, or perhaps better still, *unauthorized delegation of particular tasks*.

Thus while a weapon system may be capable of making "factual determinations" as to whether to fire or select and engage targets, these determinations are redundant to the commander's prior evaluation and judgment to use a particular weapon in accordance with IHL, the rules of engagement and all relevant treaties.[xv]

Complexity & The Limits of Human Reason

As humans are obligated under the laws of war to make determinations of discrimination, proportionality and precaution, and to meet this obligation they must have sufficient and appropriate amounts of information about a specific attack, there are necessary limits to the uses of autonomy in weapons systems.

However, there is a concern that ought to be raised with regards to how the use of autonomous weapons systems, even in seemingly permissible circumstances, may hinder the human commander's ability to make appropriate judgments about their use in a battlespace. In particular, there must be acknowledgment of how the deployment of various kinds of autonomous weapons systems, particularly swarms, creates a complex adaptive system in the battlespace, and this may adversely affect a human's ability to make appropriate judgments about discrimination, proportionality and precaution.

To explain, a complex adaptive system is a system made up of *adaptive* agents—that is agents that adapt or learn in response to their environment—that have incomplete information about their environment (bounded rationality) but they can change their strategies or behaviors based on feedback from the system or other agents. The environment (or system) is also populated by a *diverse* set of such agents. Markets, for example, are complex adaptive systems where agents try to maximize their strategies and adapt when new information comes into play.

Moreover, complex adaptive systems are not linear. This means that when one tries to estimate or predict the system's behavior, one cannot do so easily, if at all, because many agent's actions are dependent upon the actions of others. It is, what political scientists refer to as "strategic interaction."

Additionally, the system is rarely set by a single executive function (i.e., a top down governing force). The system is distributed and behavior can emerge.

Complex systems exhibit four main behaviors:

1) *Self-Organization* (such as with swarms, schools or flocks).
2) *Chaotic Behavior* (small changes in initial conditions lead to very large later changes).
3) *Fat-Tailed Behavior* (rare events occur much more often than would be predicted by a normal distribution).

4) *Adaptive Interaction* (interacting agents modify their strategies in diverse ways as experience accumulates).[xvi]

It is likely that parties to an armed conflict will create a complex adaptive system for two reasons. First, humans form part of the system and they are learning agents who adapt. Second, with more robust machine learning deployed on or with autonomous systems, the weapon systems will also be adaptive agents. Even if this is simply two parties to a conflict using swarms to attack each other, say for suppression of air defenses, they will create a micro complex adaptive system in that moment. Presently there is no way to predict with certainty what will happen when one party's swarm encounters another (adversary) party's swarm.

What is more, given that we are dealing with hierarchical organizations (states and their militaries) there is also a high probability that this complex adaptive system will generate large numbers of *loops*. A loop recirculates signals and resources. It also permits positive and negative feedback into the cycle. With increased sensors, as well as multirole applications where weapons and sensors gather and circulate information continuously, knowing if the systems are caught in a loop is not easy.

Given that analysis of complex adaptive systems is challenging, and the likelihood of creating loops, especially "tangled loops" where agents are part of multiple loops, predicting what will happen will be increasingly difficult, if not impossible. This nonlinearity yields a troubling conclusion: commanders may not actually have sufficient information to deploy AWS in a battlespace because they cannot know what the system will do when it encounters adversary systems. Simulations of adversary behavior can ameliorate some uncertainty, but they cannot provide the kind of predictability that has accompanied weapons systems reviews in the past.

The operational risk of fielding such systems may be so great that one must take into consideration not merely what one foresees, given what a commander knows about the weapon system and the state of hostilities at that moment in time, *but estimate what may even seem impossible*.[xvii]

Conclusion

The foregoing analysis offers two substantive conclusions. First, if one believes that the laws of armed conflict place obligations on human commanders, and these obligations cannot be delegated to weapons, then there are necessary limits in time and space on the deployment of such systems. Systems that are deployed for extended durations, say in more "permissive" environments, require a communications link and human authorization to use force. Short duration weapons, say single use swarm munitions or loitering munitions, would be held to strict limitations on time and space.

However, we should not be overly confident even in short duration autonomous weapons systems. As human commanders are the bearers of obligations to make decisions regarding discrimination, proportionality and precaution, they must make these decisions with all of the available information at the time, in reasonably good faith. If autonomous weapons systems are adaptive, if emergent behaviors are likely to occur, then human commanders may not possess the information needed to make such judgments, and thus it may be impermissible to field these weapon systems.

Endnotes

[i] Draft Recommendations. United Nations Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems (April 2016). Available at: http://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2016/meeting-experts-laws/documents/DraftRecommendations_15April_final.pdf

[ii] Government of Canada. "Canadian Food for Thought Paper: Context, Complexity and LAWS." (April 2016) Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/C6F73401FA55F58FC1257F850043AB3A/$file/2016_LAWS+MX_CountryPaper+Canada+FFTP2.pdf

[iii] Government of Switzerland. "Towards a 'Compliance-Based' Approach to LAWS" Informal Working Paper (March 2016). Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/D2D66A9C427958D6C1257F8700415473/$file/2016_LAWS+MX_CountryPaper+Switzerland.pdf

[iv] Government of France. "Mapping of Technological Developments" Non Paper (April 2016) Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/B9E3E8041CE4D326C1257F8F005A31E2/$file/2016_LAWSMX_CountryPaper_France+MappingofTechnicalDevelopments+EN.pdf

[v] International Committee for the Red Cross. "Autonomous Weapons Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons" (September, 2016). Available at: https://www.icrc.org/en/publication/4283-autonomous-weapons-systems

[vi] Switzerland op. cit. footnote 3.

[vii] United Nations Institute for Disarmament Research. "Statement of the UN Institute for Disarmament Research at the CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems" delivered by Ms. Kerstin Vignard (April 2016). Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/86C96CC8C7A932DCC1257F930057C0E3/$file/2016_LAWS+MX_GeneralExchange_Statements_UNIDIR.pdf

[viii] For example, Italy, Poland, and The Netherlands endorse the notion of meaningful human control, while the United Kingdom has agreed that some form of human control is required. Both the United States and Israel have accepted "appropriate human judgment" as their preferred frame. For citations see country statements at: http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2?OpenDocument

[ix] Roff, Heather M. and Richard Moyes. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons." Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, April 2016. www.article36.org/autonomous-weapons/roff-moyes-fli-paper/

[x] *Ante bellum* would include all of the design, acquisition, development, training, doctrine and "before war" mechanisms. *In bello* are all the rules, regulations, laws, and doctrine that structure how belligerents may permissibly fight. *Post Bellum*, or after the war, are all those accountability measures and institutions that come into play after conflict.

[xi] Government of Israel. "Statement on Lethal Autonomous Weapons Systems" delivered by Ms. Maya Yaron. Group of Experts Meeting on Lethal Autonomous Weapons Systems (LAWS) April 11-15 2016. http://www.unog.ch/80256EDD006B8954/(httpAssets)/5951D4CF7936ADE3C1257F9A004B62D6/$file/2016_LAWS_+MX_ChallengestoIHL_Statements_Israel.pdf

[xii] United States Department of Defense. *Department of Defense Law of War Manual* (June 2015)

[xiii] This is similar to Paul Scharre's point in his work on operational risk and AWS. The human must be the "fail safe." Scharre, Paul. "Autonomous Weapons and Operational Risk" Center for a New American Security (2016). https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk .

[xiv] Steinbruner, John. D. "Choices and Trade-Offs" in *Managing Nuclear Operations*, Eds. Ashton B. Carter, John D. Steinbruner and Charles A. Zraket (The Brookings Institution Press, 1987): 539.

[xv] United States Department of Defense Law of War Manual.

[xvi] Brownlee, Jason. "Complex Adaptive Systems" CIS Technical Report 070302A (March 2007). Holland, John. *Hidden Order: How Adaptation Builds Complexity*. (Addison Wesley Publishing, 1995). Holland, John H. *Complexity: A Very Short Introduction* (Oxford University Press, 2014). Kauffman, Stuart. "Principles of Adaptation in Complex Systems" in Lectures in the Sciences of Complexity, ed. Steinn (1989): 619-712. Lansing, J. Stephen. "Complex Adaptive Systems" *Annual Review of Anthropology*, Vol. 32 (2003): 183-204.

[xvii] Scharre, Paul. "Autonomous Weapons and Operational Risk" Center for a New American Security (2016). https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk .

Project Information:
"Artificial Intelligence, Autonomous Weapons and Meaningful Human Control"

This project seeks to generate a framework for considering how artificial intelligence in weapons systems can be subject to meaningful human control (MHC). We develop conceptual and practical elements of MHC, primarily by establishing a multidisciplinary and multi-stakeholder team to inform academic and policy relevant guidance. We attempt to identify regulative and socially beneficial values associated with the human control in the use of lethal force. To that end, the project provides conceptual and empirical value to a multi-stakeholder audience, by providing policy briefs and the first ever dataset on automated functions on presently deployed weapons systems. The dataset, the Survey of Autonomy in Weapons Systems, is freely available for download.

To download a copy of this paper and the dataset, please visit:
https://globalsecurity.asu.edu/robotics-autonomy

Additionally, the first paper in the series, "Meaningful Human Control, Artificial Intelligence, and Autonomous Weapons," outlines the content of meaningful human control and its application to autonomous weapons systems. That paper, co-authored with Richard Moyes of Article 36, can be found on Article 36's website, here:

http://www.article36.org/autonomous-weapons/roff-moyes-fli-paper/

Citation Information:

Roff, Heather M. "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons" Briefing paper prepared for the Review Conference of the Convention on Conventional Weapons, December 2016.

Funding Disclosures:

This paper is informed by a two-day workshop, funded by a generous grant awarded to Arizona State University by the Government of Canada's Defence Targeted Engagement Program. The positions expressed here are the author's, informed by discussions at the workshop, and do not represent the opinions of the Government of Canada, The University of Oxford, Arizona State University or New America.

Additionally, the first briefing paper and dataset was funded by a grant from the Future of Life Institute. Grant Number 2015-146617